The Stata Journal (2018)
18, Number 2, pp. 447–460

Inference for clustered data

Chang Hyung Lee
Department of Economics
University of California, Santa Barbara
Santa Barbara, CA
clee00@umail.ucsb.edu

Douglas G. Steigerwald
Department of Economics
University of California, Santa Barbara
Santa Barbara, CA
doug@ucsb.edu

Abstract. In this article, we introduce `clusteff`, a community-contributed command for checking the severity of cluster heterogeneity in cluster–robust analyses. Cluster heterogeneity can cause a size distortion leading to underrejection of the null hypothesis. Carter, Schnepel, and Steigerwald (2017, *Review of Economics and Statistics* 99: 698–709) develop the effective number of clusters to reflect a reduction in the degrees of freedom, thereby mirroring the distortion caused by assuming homogeneous clusters. `clusteff` generates the effective number of clusters. We provide a decision tree for cluster–robust analysis, demonstrate the use of `clusteff`, and recommend methods to minimize the size distortion.

Keywords: st0531, clusteff, cluster heterogeneity

1 Model

The basic setting is to consider a specification for n observations grouped into G clusters of the form

$$y_{ig} = \mathbf{x}_{ig}^T \boldsymbol{\beta} + u_{ig} \tag{1}$$

where observation i belongs to cluster g with n_g observations in cluster g. We assume $\mathbb{E}(u_{ig}|\mathbf{x}_{ig}) = 0$, so (1) captures the conditional mean of y_{ig}. The error term u_{ig} is allowed to have arbitrary correlation within a cluster, where $\boldsymbol{\Omega}_g$ is the covariance matrix for cluster g conditional on $\mathbf{x}_g$ but is assumed to be independent across clusters. In this article, we provide a command that estimates the effective number of clusters, which is a diagnostic tool used to measure severity of cluster heterogeneity (including lack of balance in the covariate matrix) derived by Carter, Schnepel, and Steigerwald (2017).

The question of interest is to test the null hypothesis $H_0 : \mathbf{a}^T\boldsymbol{\beta} = \mathbf{a}^T\boldsymbol{\beta}_0$, where $\boldsymbol{\beta}_0$ is the value of $\boldsymbol{\beta}$ under the null hypothesis and $\mathbf{a}$ is a vector selecting the coefficients to be included in the test. We focus on the conventional test statistic constructed from $\widehat{\boldsymbol{\beta}}$—the ordinary least-squares (OLS) estimator of $\boldsymbol{\beta}$ in (1),

$$t = \frac{\mathbf{a}^T\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)}{\sqrt{\mathbf{a}^T\widehat{\mathbf{V}}\mathbf{a}}} \tag{2}$$

where $\widehat{\mathbf{V}}$ is a cluster–robust estimator of $\mathbf{V}$—the variance of $\widehat{\boldsymbol{\beta}}$ conditional on the covariate matrix $\mathbf{X}$. The cluster–robust estimator of $\mathbf{V}$ is

$$\widehat{\mathbf{V}} = c\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g^T\widehat{\mathbf{u}}_g\widehat{\mathbf{u}}_g^{\mathrm{T}}\mathbf{X}_g\right)\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

where $\mathbf{X}_g$ and $\mathbf{u}_g$ are the covariate matrix and error, respectively, for cluster g and $c = \{G(n-1)\}/\{(G-1)(n-k)\}$ is designed to partially offset the downward bias in $\widehat{\mathbf{V}}$.

The consistency of $\widehat{\mathbf{V}}$ and the asymptotic normality of t is established under general conditions in Carter, Schnepel, and Steigerwald (2017). As they describe, consistency of $\widehat{\mathbf{V}}$ cannot be established simply by allowing the number of observations n to grow without bound but rather depends crucially on allowing the number of clusters G to grow without bound. To understand why this is so, consider a dataset with a fixed number of clusters but an increasing number of observations in each cluster. As more observations are added to each cluster, the dimension of $\widehat{\mathbf{u}}_g$ grows and more parameters are added to $\boldsymbol{\Omega}_g$. Consequently, $\widehat{\mathbf{u}}_g\widehat{\mathbf{u}}_g^{\mathrm{T}} := \widehat{\boldsymbol{\Omega}}_g$ is not a consistent estimator of $\boldsymbol{\Omega}_g$, and consistency of $\widehat{\mathbf{V}}$ can be obtained only by averaging $\widehat{\boldsymbol{\Omega}}_g$ over an increasingly large number of clusters. Thus, the size of G is often advocated as a guide to inference. According to this guide, if G is large (say, greater than 50), then the appropriate critical values to use when assessing t are obtained from a normal distribution.

The standard practice of using G as the sole criterion when selecting critical values relies on an assumption that clusters are homogeneous in the sense that $\mathbb{E}\left(\mathbf{X}_g^{\mathrm{T}}\boldsymbol{\Omega}_g\mathbf{X}_g\right)$ is identical over clusters. A sufficient condition for this assumption is that all clusters have identical size, $n_g = n/G$; covariate matrices, $\mathbf{X}_g$, that are identical over g; and covariance matrices, $\boldsymbol{\Omega}_g$, that are identical over g. Because these sufficient conditions rarely occur in practice, Carter, Schnepel, and Steigerwald (2017) investigate the behavior of t when clusters are heterogeneous. They find that the test often falsely rejects (that is, the critical values from a normal distribution are too small) under cluster heterogeneity.

Importantly, Carter, Schnepel, and Steigerwald (2017) report a simple measure that can detect the extent to which cluster heterogeneity affects the test statistic. The measure adjusts the number of clusters downward to reflect the degree of cluster heterogeneity such that the larger amounts of cluster heterogeneity correspond to greater downward adjustment in the number of clusters. The resultant adjusted measure is the effective number of clusters. If the effective number of clusters is small regardless of the magnitude of G, critical values that are larger than those from a normal distribution should be used. These critical values may be obtained from a Student's t distribution or from bootstrapping, as explained below.

Observe that $\mathbf{V} = (\mathbf{X}^T\mathbf{X})^{-1}\sum_g(\mathbf{X}_g^T\boldsymbol{\Omega}_g\mathbf{X}_g)(\mathbf{X}^T\mathbf{X})^{-1}$ with $\gamma_g = \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}_g^T\boldsymbol{\Omega}_g\mathbf{X}_g)(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}$. Thus $\mathbf{a}^T\mathbf{V}\mathbf{a} = \sum_g\gamma_g$. Following Carter, Schnepel, and Steigerwald (2017), we denote the effective number of clusters as G^* and define it as

$$G^* = \frac{G}{1+\Gamma} \qquad \Gamma = \frac{1}{G}\sum_{g=1}^{G}\left(\frac{\gamma_g - \overline{\gamma}}{\overline{\gamma}}\right)^2$$

with $\overline{\gamma} = G^{-1}\sum_g \gamma_g$. Simply put, cluster homogeneity requires $\gamma_g = \gamma$ for all clusters, so variation in γ_g arises from cluster heterogeneity. If the clusters are homogeneous, then $\Gamma = 0$ and $G^* = G$. If the clusters are heterogeneous, then $\Gamma > 0$ and $G^* < G$. A greater difference between G^* and G is indicative of more heterogeneous clusters.

Here special attention is required to $\mathbf{a}$, which is a selection vector of length k. The selection vector is derived from the hypothesis to be tested, $H_0 : \mathbf{a}^T\boldsymbol{\beta} = \mathbf{a}^T\boldsymbol{\beta}_0$. Consequently, a unique value of G^* is generated based on each hypothesis to be tested. To be clear, the method is appropriate for tests of hypotheses on single coefficients, for example, $H_0 : \beta_1 = 0$, as well as for linear combination of coefficients, $H_0 : \beta_1 + \beta_2 = 0$.

If G^* is small, inference should be undertaken with care. Carter, Schnepel, and Steigerwald (2017) show that t is asymptotically normal as $G^* \to \infty$, which means the normal approximation should work well if G^* is large. If G^* is small, then the appropriate critical values are larger than those from a normal distribution, and mistakenly applying the normal critical values leads to incorrectly rejecting the null hypothesis far too often (the empirical size of the test exceeds the nominal size of the test). They find that the empirical size of a test remains close to the nominal size using normal critical values for G^* greater than 25.

In practice, G^* must be estimated because it is a function of the unknown within-cluster error covariance matrix $\boldsymbol{\Omega}$. Unfortunately, we cannot use the residuals to estimate G^*, because use of the residuals to construct both the critical values and the test statistic induces pretest bias. Rather, G^* is estimated by G^{*A}, which is constructed under the assumption of perfect within-cluster error correlation.[1] Because increasing the within-cluster correlation tends to increase cluster heterogeneity, the estimate G^{*A} is designed to guard against this "worst-case scenario" in which the errors are perfectly correlated within clusters.

We recommend estimating G^* as a first step in testing a model with a clustered error structure to credibly rule out size distortion from cluster heterogeneity. Application of the effective number of clusters need not be limited to small to moderate G, because a large G does not guarantee G^* to be large under cluster heterogeneity. Carter, Schnepel, and Steigerwald (2017) demonstrate the fallibility of assuming large G^* based on large G using the dataset clustered at the industry level from Hersch (1998). The dataset contains 5,960 observations in 211 clusters. Conventional wisdom suggests that the number of clusters in this case is large enough to assume an approximately normal distribution for the test statistic. However, calculating the effective number of clusters reveals that the dataset suffers from severe cluster heterogeneity with $G^{*A} = 19$, and the normal critical values are likely too small. In essence, variation in the covariate matrix across clusters yields substantial variability in the estimator of the standard error that appears in the denominator of the test statistic. Accounting for this variability

1. The estimation procedure for G^{*A} used by the program is further discussed in the next section.

enlarges the critical values. We also note that in applications where the key question of interest involves the response to treatment in specific clusters, the key criterion is not the overall value of G^{*A} but rather the effective number of treated clusters (and the effective number of control clusters).

In section 2, we detail the command. In section 3, we follow with a decision tree for selecting the appropriate method of inference. In section 4, we present an example on use of the decision tree.

2 The clusteff command

2.1 Syntax

`clusteff` *varlist* [*if*] [*in*], `cluster(`*varname*`)` [`test(`*varname*`)` `selection(`*string*`)` `noconstant` `covariance(`*real*`)`]

2.2 Description

`clusteff` estimates the effective number of clusters (G^*) devised by Carter, Schnepel, and Steigerwald (2017) using a vector of independent variables, a clustering variable, and a selection vector. The command uses *varlist* as a list of variables to be included in the estimation procedure with the data clustered by the variable specified in the `cluster()` option and the hypothesis test of interest defined by either the `selection()` option or the `test()` option.

2.3 Options

`cluster(`*varname*`)` states the clustering variable. `cluster()` is required.

`test(`*varname*`)` specifies a selection vector if the null hypothesis of interest involves a single covariate. Suppose a user aims to test the null hypothesis, $H_0 : \beta_2 = 0$, using a linear model of the following form: $y = \beta_0 + \beta_1 x + \beta_2 z + u$. Then,

```
clusteff x z, cluster(clustervar) test(z)
```

generates the relevant effective number of clusters.

`selection(`*string*`)` allows users to define their own selection vector. The *string* is a vector of values selecting the coefficients to be tested corresponding to the vector $\mathbf{a}$ in the null hypothesis, $H_0 : \mathbf{a}^T\boldsymbol{\beta} = 0$. The order of covariates in *varlist* must match the order of elements in the selection vector. This option is especially useful if the null hypothesis of interest involves more than one covariate. For example, if a user is testing the null, $H_0 : \beta_1 + \beta_2 = 0$, stating

```
clusteff x z, cluster(clustervar) selection(1 1)
```

estimates the appropriate effective number of clusters.

The `test()` and `selection()` options may not be specified simultaneously. If both the `selection()` and `test()` options are specified, the `selection()` option overrides the `test()` option by generating a selection vector based on the `selection()` option while ignoring the `test()` option.

The number of elements in a selection vector may not exceed the number of variables. However, the number of specified elements in a selection vector is allowed to be smaller than the number of variables. The program fills empty elements with zeros such that `selection(1 0)` or `selection(1)` generates the effective number of coefficients under the null hypothesis, $H_0 : \beta_1 = 0$.

If a user omits both the `test()` and `selection()` options, the program estimates an effective number of clusters under an assumption that the first variable in *varlist* is the covariate of interest. In the above example, omitting both of the options is equivalent to specifying `test(x)`, `selection(1)`, or `selection(1 0)`.

`noconstant` determines whether a linear model to be tested contains a vector of constants. If this option is specified, the program estimates an effective number of clusters without a vector of constants. Use this option when testing a linear model whose intercept is restricted at zero.

`covariance(`*real*`)` allows users to specify any real number between zero and one as the within-cluster covariance of the error used to estimate the effective number of clusters. The default is `covariance(1)`.[2] A covariance of less than one estimates a less conservative effective number of clusters relative to the default in which perfect within-cluster error correlation is imposed.

2.4 Estimation procedure

Generating a true value of an effective number of clusters (G^*) requires the underlying error structure, $\mathbb{E}(\mathbf{u}_g \mathbf{u}_g^T)$, to be known. However, using residuals from a regression $\widehat{\mathbf{u}}_g$ to construct critical values renders a test invalid (Carter, Schnepel, and Steigerwald 2017). Instead, Carter, Schnepel, and Steigerwald (2017) suggest using a 1-by-n_g vector of ones, $\boldsymbol{\iota}_g$, in place of $\mathbf{u}_g$ to impose a perfect within-cluster error correlation as a conservative approach. `clusteff` uses the above estimation procedure to generate an estimate of G^*, G^{*A}, as outlined below.

$$G^{*A} = \frac{G}{1 + \Gamma^A}$$

2. The program limits the maximum covariance at 0.9999 instead of 1 because of limits on floating-value precision in Mata. This produces a more stable estimator.

where

$$\Gamma^A = \frac{1}{G}\sum_{g=1}^{G}\left(\frac{\gamma_g^A - \overline{\gamma}^A}{\overline{\gamma}^A}\right)^2$$

and

$$\gamma_g^A = \mathbf{a}^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}_g^T\boldsymbol{\iota}_g\boldsymbol{\iota}_g^T\mathbf{X}_g\right)\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{a}$$

Any valid input in `selection(`*string*`)` or `test(`*string*`)` is converted to a selection vector, $\mathbf{a}$, used to generate G^{*A}. The program performs a matrix multiplication estimating a scalar value of G^{*A}.

3 Decision tree

What is the correct approach for a practitioner with clustered data? As noted above, a key quantity in determining the best method of inference is the effective number of clusters. Thus, the decision begins with an estimate of this quantity for a given sample. If the estimated effective number of clusters, G^{*A}, is at least 50, then one should use the statistic (2) with critical values from a normal distribution. If G^{*A} is less than 50, then a leading approach would be to use (2) but with critical values obtained in a different way. Cameron, Gelbach, and Miller (2008) and MacKinnon and Webb (2017) find that the wild bootstrap, which delivers critical values that are larger than those from a normal distribution, brings the empirical size of the test much closer to the nominal size.

Note that for models where the coefficient of interest is a cluster-level treatment, G^{*A} should be calculated separately for both the treated clusters and the control clusters. If either of these measures of G^{*A} is less than 25, even if the overall effective number of clusters exceeds 50, then again the wild bootstrap could be used to obtain more accurate critical values.[3]

The wild bootstrap begins by drawing, with replacement, from the collection of cluster residual vectors $\{\widehat{\mathbf{u}}_g\}_{g=1}^{G}$. Each residual vector is multiplied by either 1 or -1 with equal probability. Then, the resultant residual vectors are combined with the observed regressors to produce bootstrap samples. Complete details are provided in Cameron, Gelbach, and Miller (2008); Cameron and Miller (2015); and MacKinnon and Webb (2017). Community-contributed commands `cgmwildboot` by Caskey (2010) and `boottest` by Roodman (2015) can be used to generate p-values via wild bootstrap.

3. With clusters identical to the size of U.S. states, MacKinnon and Webb (2017) show that severe underrejection can occur if there are fewer than seven treated or untreated clusters. Ferman and Pinto (2015) study the case of a small number of treated clusters in a difference-in-differences setting.

For datasets that have a small effective number of clusters, either overall or within the treatment group (while rare, a similar issue arises if the control group has a small effective number of clusters), there are alternatives to the wild bootstrap. If interest centers on the coefficient of a covariate that varies within clusters and there are a large number of observations in each cluster, then Ibragimov and Müller (2010) propose an alternative test statistic. To illustrate their method, we first rewrite (1) to distinguish an observation-level covariate, x_{ig}, from a cluster-level covariate, z_g,

$$y_{ig} = \alpha + \beta x_{ig} + \delta z_g + u_{ig} \tag{3}$$

The test statistic is derived by first estimating $\widehat{\beta}_g$ separately for each cluster. Note that α and δ are both absorbed in the cluster level intercept and so are not separately identified. The test statistic is

$$t_{\text{IM}} = \frac{\sqrt{G}\left(\overline{\widehat{\beta}} - \beta\right)}{s_{\widehat{\beta}}}$$

where $\overline{\widehat{\beta}} = 1/G \sum_{g=1}^{G} \widehat{\beta}_g$ and $s^2 = 1/(G-1) \sum_{g=1}^{G} (\widehat{\beta}_g - \overline{\widehat{\beta}})^2$. Under the cluster assumption, $\widehat{\beta}_g$ is independent of $\widehat{\beta}_h$, and if n_g is sufficiently large, then $\widehat{\beta}_g$ has a normal asymptotic null distribution with mean β and variance σ_g^2. Of course, if $\widehat{\beta}_g$ is a normal random variable and $\sigma_g^2 = \sigma^2$, then $t_{\text{IM}} \sim t(G-1)$. One would think that allowing σ_g^2 to vary would result in a test statistic with larger critical values than those from the Student's $t\,(G-1)$. Surprisingly, for a test with a nominal size of 5%, the critical values for t_{IM} are smaller than the critical values from a Student's $t\,(G-1)$. Thus, combining t_{IM} with the critical values from a $t\,(G-1)$ yields a test whose size will not exceed the nominal size of 5%. Note that such a result does not hold for a test with a nominal size of 10%, so selection of a nominal size of 5% is important. In comparing this method with the wild bootstrap, Ibragimov and Müller (2016) find that t_{IM} is better at eliminating the size distortion for a very small number of heterogeneous clusters with large n_g.

If interest centers on the coefficient of a covariate that does not vary within clusters, and n_g is large, then Donald and Lang (2007) propose an alternative test statistic. To illustrate their method, we begin with the regression (3), where the error has an error-components structure

$$u_{ig} = \rho_g + \epsilon_{ig}$$

The first step is to construct the OLS fixed-effects estimator from

$$y_{ig} = \beta x_{ig} + c_g + \epsilon_{ig}$$

yielding $\{\widehat{c}_g\}_{g=1}^{G}$. The second step is to construct the OLS estimator of δ from

$$\widehat{c}_g = a + \delta z_g + v_g$$

yielding $\widehat{\delta}$. For the $H_0 \colon \delta = \delta_0$, the test statistic is

$$t_{\text{DL}} = \frac{\left(\widehat{\delta} - \delta_0\right)}{s_{\widehat{\delta}}}$$

where $s^2_{\widehat{\delta}} = s^2/\{\sum_{g=1}^{G}(z_g - \overline{z})^2\}$ and $s^2 = 1/(G-2)\sum_{g=1}^{G}(\widehat{v}_g^2)$. The distribution of t_{DL} is approximately Student's $t\,(G-2)$, so again the critical values are larger than those from a normal distribution.

There are two caveats to using this test statistic. The first is that, as in the case of t_{IM}, the number of observations in each cluster must be large. The second is that the distribution of the test statistic depends crucially on homogeneity across clusters (in essence, n_g and $\overline{x}_g$ are both identical across clusters). Thus, if G^{*A} differs substantially from G, indicating that these homogeneity conditions do not hold, then it may not be appropriate to use t_{DL}.

MacKinnon and Webb (2017) investigate the relative performance of the wild bootstrap and t_{DL}. For data in which each cluster has 40 observations, but varying covariates across clusters, the wild bootstrap and t_{DL} can have comparable empirical size. Importantly, the comparable size requires the use of G^{*A} rather than G when constructing the critical values from a Student's t distribution. In other words, if t_{DL} is used with critical values from the $t(G-2)$ distribution, then the wild bootstrap outperforms it in the sense of more accurate size. A second set of simulations allows the cluster sizes to vary together with varying covariates across clusters. In these models with more pronounced cluster heterogeneity, the wild bootstrap outperforms t_{DL} and delivers the most accurate size.

In figure 1, we provide a decision tree that encapsulates this discussion.

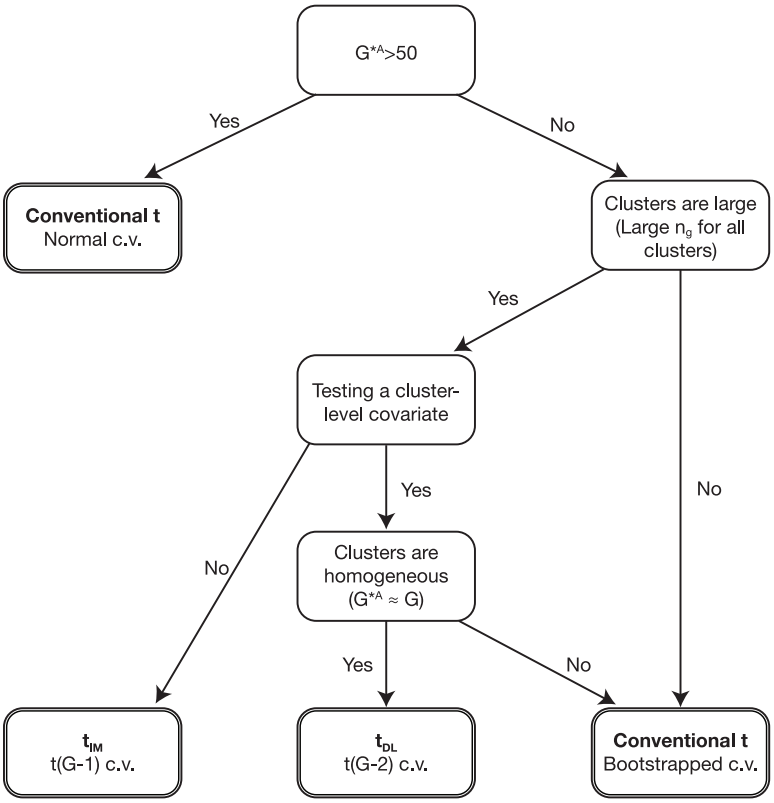

Figure 1. Decision tree

4 Example

We recommend using `clusteff` as a simple check to verify validity of analyses and to find an optimal method to use to minimize both the amount of computational power required and the size distortion. This section uses an example from the economics literature to demonstrate the use of `clusteff` in analyses of clustered samples.

4.1 Clustering at the state level

Voena (2015) studies changes in the employment decisions of married women that result from the introduction of unilateral divorce laws. The introduction of unilateral divorce, under which divorce can be initiated without mutual consent of both partners, increases the probability of divorce. If women have fewer resources in divorce than in marriage, they may need to insure themselves against this potential loss of resources by working

while married (thereby building their human capital). Because states have different rules governing the distribution of property upon divorce, the strength of this effect is likely to vary across states. In states with "equitable distribution", under which women often have fewer resources after divorce, this effect is likely to be most pronounced. In states with community property, under which each partner gets an equal share of the resources, this effect is likely to be weaker. Female labor market participation, therefore, is likely to be more responsive to the divorce law reform in states with "equitable property" division.

To test the theory, we fit a linear probability model for the labor force participation by women in household i, state s, and year t. Key coefficients of interest are on the interaction covariates, which are indicators for whether state s has unilateral divorce and (say) community property in year t. The corresponding component of the regression model is

$$\beta_1 \left(\{\text{uni}_{st}\} \times \{\text{comprop}_{st}\}\right) + \beta_2 \left(\{\text{uni}_{st}\} \times \{\text{eqdistr}_{st}\}\right)$$

where $\{\text{uni}_{st}\}$ takes the value 1 if unilateral divorce is legal in state s in year t, $\{\text{comprop}_{st}\}$ takes the value 1 if community property rules are used to govern divorce, and $\{\text{eqdistr}_{st}\}$ takes the value 1 if equitable distribution rules are used to govern divorce. The individual hypotheses being tested are $H_0 : \beta_i = 0$, where $i = 1, 2$.

The conventional cluster–robust t statistic (2) is estimated where clustering is at the state level. The number of clusters is 51, corresponding to the 50 states and the District of Columbia. The number of observations from each state varies widely, from 3 to 3,552. This large variation in cluster size indicates substantial cluster heterogeneity. As an initial indicator, we compute the effective number of clusters accounting only for variation in cluster sizes (that is, ignoring how the covariates change over clusters).[4] Such a calculation provides a quick indicator of the degree of cluster heterogeneity. For this dataset, $G^{*A} = 13$, well below the cutoff for Gaussian inference. As noted above, this approximation of G^* is likely to be conservative because it is based on an intracluster correlation of 1. An alternative approximation, which assumes no intracluster correlation and so is much less conservative, can be constructed by replacing the unit matrix in γ_g^A with the identity matrix. For this dataset, this less conservative approximation yields $G^{*A} = 20$, again below the cutoff for Gaussian inference. All initial evidence points to the need to move away from the use of critical values from the normal distribution.

Because the form of the conditional expectation function is not known, Voena (2015, table 2, columns 5–8, 2,314) provides four regression approximations that differ in the number of controls. In the following table, we present the OLS estimate and the cluster–robust standard error reported by Voena, followed by the effective number of clusters and the bootstrapped confidence interval (CI) in brackets.

4. This computation corresponds to a test on the intercept.

Table 1. Replication results

Variables	(1) Employed	(2) Employed	(3) Employed	(4) Employed
uni × comprop	−0.0377	−0.0389	−0.0575	−0.0488
	(0.0164)	(0.0175)	(0.0175)	(0.0177)
G^{*A}	1.9191	1.9227	4.9457	5.0394
Bootstrapped 95% CI	[−0.0868, 0.0056]	[−0.1096, 0.0073]	[−0.1205, −0.0204]	[−0.1181, −0.0092]
uni × eqdistr	−0.0279	−0.0263	−0.0265	−0.0298
	(0.0306)	(0.0314)	(0.0387)	(0.0414)
G^{*A}	4.9574	4.9630	13.3717	12.8005
Bootstrapped 95% CI	[−0.1089, 0.0372]	[−0.1018, 0.0360]	[−0.1235, 0.0553]	[−0.1228, 0.0541]
Year fixed effects	Yes	Yes	Yes	Yes
Age dummies	Yes	Yes	Yes	Yes
Children dummies	No	Yes	Yes	Yes
State fixed effects	No	No	Yes	Yes
Polyn yrs. married	No	No	No	Yes
Observations	44,808	44,808	44,808	39,824
Individual fixed effects	3,437	3,437	3,437	2,607

Note: Replication of columns 5–8 from table 2 of Voena (2015). Standard errors are clustered at the state level, and critical values are generated by the wild bootstrap procedure with 1,000 replications. The third row estimates the effective number of clusters, while the fourth row presents the wild bootstrap CI between 2.5 and 97.5 percentiles. Although standard errors reported here are generated using data and codes provided by Voena (2015), they slightly differ from table 2 of Voena (2015). However, the size of the difference does not change the inference significantly.

For each of the null hypotheses under test, the effective number of clusters is obtained within Stata using `clusteff`. For example, consider the test of β_1 in column 1, for which the command is

```
. clusteff uni_comprop uni_title uni_eqdistr comprop eqdistr d_age*
> yrd* i.person, cluster(state) test(uni_comprop)
```

We list all covariates included in the model in *varlist*, specify `state` as the clustering variable, and include the null hypothesis to be tested. A portion of the output is

```
Number of clusters: 51
Estimated effective number of clusters: 1.919089
Warning: G* estimated to be below 50.
```

where the effective number of clusters corresponds to the coefficient being tested.

With such a small value for G^{*A}, and such substantial cluster heterogeneity, there are two potential methods of inference from the decision tree. The first is to combine the standard test statistic t with critical values obtained from the wild bootstrap. A second possibility, appropriate for regressors that vary within states, is to use t_{IM} with critical values from the Student's $t(50)$ distribution. To construct t_{IM}, we must be able to estimate β_1 and β_2 for each state separately. Yet, for some states, $\{\text{uni}_{st}\} \times \{\text{comprop}_{st}\}$

is always 0, rendering β_1 unidentified for these states.[5] Hence, we report wild bootstrap critical values for t below the approximations of G^* in table 1.

We use `boottest`, the aforementioned community-contributed command for Stata, to obtain the wild bootstrap critical values. The first line of the code runs a regression, and the second line of the code performs wild bootstrap to generate critical values for the specified null.

```
regress participation uni_comprop uni_title uni_eqdistr comprop  ///
    eqdistr d_age* yrd* i.person chd*, cluster(state)
boottest uni_comprop=0 uni_eqdistr=0
```

In this case, using wild bootstrap instead of conventional t critical values yields a wider CI. The difference in CI is less relevant for the estimated coefficients on uni$\times$eqdistr because they remain insignificant regardless of the method used to infer the significance. On the other hand, the estimated coefficients on uni $\times$ comprop, which are significant at the 5% level under conventional t critical values, lose significance in two of the four columns with bootstrapped CI. The change in significance suggests that alternative methods, such as wild bootstrap, are necessary when drawing inference from a dataset with a small effective number of clusters.

5 References

Cameron, A. C., J. B. Gelbach, and D. L. Miller. 2008. Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics* 90: 414–427.

Cameron, A. C., and D. L. Miller. 2015. A practitioner's guide to cluster–robust inference. *Journal of Human Resources* 50: 317–372.

Carter, A. V., K. T. Schnepel, and D. G. Steigerwald. 2017. Asymptotic behavior of a t-test robust to cluster heterogeneity. *Review of Economics and Statistics* 99: 698–709.

Caskey, J. 2010. cgm: Implementation of multi-way clustered standard errors as in Cameron, Gelbach, and Miller (2006). https://sites.google.com/site/judsoncaskey/data/.

Donald, S. G., and K. Lang. 2007. Inference with difference-in-differences and other panel data. *Review of Economics and Statistics* 89: 221–233.

Ferman, B., and C. Pinto. 2015. Inference in differences-in-differences with few treated groups and heteroskedasticity. MPRA Paper 81988, University Library of Munich. https://mpra.ub.uni-muenchen.de/81988/1/MPRA_paper_81988.pdf.

Hersch, J. 1998. Compensating differentials for gender-specific job injury risks. *American Economic Review* 88: 598–607.

5. We provide the code to construct t_{IM}, the Ibragimov–Müller (IM) t statistic in the appendix.

Ibragimov, R., and U. K. Müller. 2010. t-statistic based correlation and heterogeneity robust inference. *Journal of Business and Economic Statistics* 28: 453–468.

———. 2016. Inference with few heterogeneous clusters. *Review of Economics and Statistics* 98: 83–96.

MacKinnon, J. G., and M. D. Webb. 2017. Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics* 32: 233–254.

Roodman, D. 2015. boottest: Stata module to provide fast execution of the wild bootstrap with null imposed. Statistical Software Components S458121, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s458121.html.

Voena, A. 2015. Yours, mine, and ours: Do divorce laws affect the intertemporal behavior of married couples? *American Economic Review* 105: 2295–2332.

About the authors

Chang Hyung Lee is a PhD candidate in economics at the University of California, Santa Barbara.

Douglas G. Steigerwald joined the faculty of the Department of Economics at the University of California, Santa Barbara, after completing an MA in statistics and a PhD in economics at the University of California, Berkeley.

A Appendix

A.1 Ibragimov and Müller

Although the test using the Ibragimov and Müller (IM) test statistic is unlikely to be valid, we show how to derive the IM test statistic, t_{IM}, to demonstrate implementation of t_{IM} using Stata.

As discussed in section 3, t_{IM} is derived by calculating the coefficient of interest individually and then assuming the derived coefficients to be approximately t distributed with $G-1$ degrees of freedom. Note that this exercise does not have an analytical power, because the covariates of interest vary in some but not all clusters.[6] As far as we are aware, there is no Stata code for IM-type analysis. It is, however, fairly simple to implement in Stata without a dedicated program.

First, we keep states with within-cluster variation in the covariate of interest uni × comprop, define the cluster variable `clustvar`, and find the number of clusters (denoted `maxclustvar` here):

```
. use psid_women.dta
. egen mean_uni_comprop = mean(uni_comprop), by(state)
```

6. Only five states had adopted unilateral divorce law and community property regime in the data. Thus, all states without any variation in the interaction term must be eliminated to estimate t_{IM} for β_1.

```
. drop if mean_uni_comprop==0 | mean_uni_comprop==1
(39,126 observations deleted)
. egen clustvar = group(state)
. sort clustvar
. local maxclustvar = clustvar[_N]
```

Next, we use a loop to calculate the coefficients individually for each group, store the results, and calculate t_{IM} using the dataset from Voena (2015).

```
. generate bhat = .
(7,006 missing values generated)
. forvalues i= 1(1)`maxclustvar´ {
  2. quietly regress participation uni_comprop comprop uni_title uni_eqdistr
> eqdistr d_age* yrd* i.person if clustvar==`i´
  3. quietly replace bhat = _b[uni_comprop] if clustvar==`i´
  4. }
. collapse bhat, by(clustvar)
. quietly summarize bhat
. local t_im = r(mean)/(r(sd)/sqrt(r(N)))
. display "Mean of betahat is " r(mean)
Mean of betahat is -.31166391
. display "Standard error of betahat is " r(sd)/sqrt(r(N))
Standard error of betahat is .41951136
. display "Test statistic is " `t_im´ " distributed t with " r(N)-1
> " degrees of freedom."
Test statistic is -.74292127 distributed t with 4 degrees of freedom.
```